■ อ.ยืน ภู่วรวรรณ

ไบโออินฟอร์มาติกส์

ไบโออินฟอร์มาติกกับรหัสของสิ่งมีชีวิตที่เป็นประโยชน์ต่อการดำรงเผ่าพันธุ์ของมนุษย์

ในสมัยสงครามโลกครั้งที่สอง การสื่อสารเป็นเรื่องสำคัญของการสั่งการและประสานงานการรบ ฝ่ายสัมพันธมิตรได้ทำการค้นคว้าวิจัยและหาทางถอดรหัสลับที่ทางกองทัพเยอรมันใช้ติดต่อกัน

รหัสลับที่ส่งเป็นรหัสตัวหนังสือผ่านระบบโทรเลข หรือเป็นรหัสมอร์สที่แทนตัวอักษร แต่มีลำดับเรียงกันไป เยอรมันเข้ารหัสด้วยเทคนิคต่าง ๆ การถอดรหัสสายตัวอักขระเพื่อแยกแยะว่าข้อความที่เยอรมันส่งมีข้อความอย่างไร

อลัน ทัวริ่ง (Alan Turing) เป็นบุคคลหนึ่งที่ศาสตร์วิชาทางคอมพิวเตอร์ยกย่องมาก อลัน ทัวริ่ง เสนอวิธีคิดการคำนวณแบบเครื่องจักรลำดับ (Turing Machine) เพื่อใช้ในการถอดรหัสข้อความของเยอรมัน ทฤษฎีเครื่องจักรลำดับเป็นการคำนวณแบบลำดับจากข้อมูลที่เป็นสายวลี (String data) หลักการดังกล่าวนี้เป็นศาสตร์พื้นฐานที่ถือได้ว่าเป็นแนวคิดที่สำคัญและพัฒนาเป็นรากฐานที่สำคัญของคอมพิวเตอร์ในยุคต่อมา

ปัจจุบันพัฒนาการทางคอมพิวเตอร์เจริญก้าวหน้าไปไกลมาก ขีดความสามารถเชิงการคำนวณทำได้มาก ซีพียูยุคปัจจุบันสามารถคำนวณตัวเลขได้หลายร้อยล้านครั้งในเวลาหนึ่งวินาที จากประสิทธิภาพที่ดีขึ้นมากของคอมพิวเตอร์ ประจวบกับการสร้างเครื่องมือเพื่อศึกษาสิ่งมีชีวิตในระดับโมเลกุลได้มากขึ้น การผสมผสานศาสตร์ทางด้านคอมพิวเตอร์กับชีววิทยาจึงเกิดขึ้น โดยเฉพาะอย่างยิ่งการถอดรหัสพันธุกรรม ซึ่งมีข้อมูลทางพันธุกรรมมากมายที่จะต้องถอดรหัส และต้องใช้ขีดความสามารถเชิงการคำนวณและเทคนิคที่สูงตามด้วย ข้อมูลทางพันธุกรรมและขยายการวิธีการที่เกี่ยวกับการประมวลผลจึงเป็นที่มาของศาสตร์แขนงใหม่ที่กำลังได้รับความสนใจ และจะเป็นประโยชน์ต่อชาวโลกมหาศาลคือ ไบโออินฟอร์มาติกส์ (Bioinformatics)

สายรหัสพันธุกรรมของมนุษย์ (human genome) เป็นสิ่งที่ท้าทายความสามารถมาก เพราะประกอบด้วยหน่วยย่อยทางเคมีระดับโมเลกุล ที่เรียงต่อกันยาวถึง 3 พันล้าน ลองนึกดูว่าเรามีรหัสข้อมูลขนาด 3 พันล้านหน่วยที่ต้องถอดรหัสด้วยคอมพิวเตอร์ งานระดับนี้จึงเกี่ยวโยงกับศาสตร์คอมพิวเตอร์มากมายที่จะต้องพัฒนาตาม

การถอดรหัสพันธุกรรมเป็นเรื่องที่สำคัญและจะให้ประโยชน์แก่มวลมนุษย์ชาติมหาศาล อีกทั้งมีมูลค่าทางการค้ามาก เพราะรหัสพันธุกรรมที่ได้เกี่ยวข้องกับการรู้ถึงการเกิดโรค เกี่ยวพันกับการสร้างยารักษาโรค การปรับปรุงพันธุ์ การโคลนนิ่ง ตลอดจนการใช้ประโยชน์ในเรื่องการสร้างอาหาร และปัจจัยสี่อื่น ๆ

เมื่อความสำคัญของไบโออินฟอร์เมติกส์มีมาก อุตสาหกรรมและองค์กรเอกชนหลายแห่ง รวมถึงองค์กรภาครัฐต่าง ๆ ที่ไม่มุ่งการค้าได้ทำการศึกษาวิจัย ข้อมูลทางพันธุกรรมนี้จะมีค่ามากมาย เชื่อกันว่า ภายในปี 2005 มูลค่าของข้อมูลในตลาดที่ทำการค้าซื้อขายข้อมูลทางพันธุกรรมนี้จะมีค่ากว่า 2 พันล้านเหรียญสหรัฐฯ

กลุ่มที่ทำการศึกษาและทำการเก็บข้อมูลสายพันธุกรรมมนุษย์มีสองกลุ่ม และได้ดำเนินการมาแล้วกว่าสิบปี กลุ่มแรกคือ โครงการฮิวแมนจีโนม (Human Genome Project) http://www.ornl.gov/hgmis/ ส่วนกลุ่มที่สองคือ บริษัท ซีลีรา จีโนมิกส์ Celera Genomics http://www.celera.com

โครงการฮิวแมนจีโนม เป็นโครงการระหว่างประเทศที่ร่วมมือกันทำงานหลายองค์กร โดยใช้เงินลงทุนรวมเป็นเงินกว่าสามพันล้านเหรียญ ในตลอดระยะเวลาสิบห้าปีที่ผ่านมาโดยช่วยกันทำการจัดทำข้อมูลสายรหัสพันธุกรรม และรวบรวมเก็บเป็นฐานข้อมูลกลางที่เรียกเข้าถึงหรือดาวน์โหลดมาปรับปรุงข้อมูลและใช้ร่วมกัน การเก็บฐานข้อมูลจัดดำเนินการโดย National Center for Biotechnology Information (ที่สหรัฐอเมริกา) สำหรับที่ยุโรปดำเนินการเก็บรวบรวมข้อมูลที่ European Bioinformatics Institute ส่วนที่ญี่ปุ่นเก็บไว้ที่ DNA Data Bank สำหรับบริษัท ซีลีรา มีฐานข้อมูลที่จัดทำขึ้นโดยบริษัท ซึ่งบริษัทเป็นผู้รับทำข้อมูลให้กับบริษัทผลิตยา และบริษัทที่ทำธุรกิจทางด้านไบโอเทคโนโลยีอื่น ๆ อีก

ที่สำคัญยิ่งและได้รับการจดบันทึกไว้ในประวัติศาสตร์ คือ ราวเดือนพฤษภาคม ปี ค.ศ. 2000 บริษัท ซีลีรา ประกาศความสำเร็จในการสร้างฐานข้อมูลรหัสพันธุกรรมของมนุษย์ได้สำเร็จ ซึ่งเร็วกว่าที่คาดไว้ ทั้งนี้เพราะระบบซูเปอร์คอมพิวเตอร์ที่บริษัทใช้ทำงานได้ดีกว่าที่ตั้งเป้าหมายไว้มาก โดยบริษัทจะจัดทำข้อมูลได้สำเร็จสมบูรณ์ และสามารถเผยแพร่ได้ในปี ค.ศ. 2001 นี้ ส่วนของโครงการฮิวแมนจีโนมจะได้เวอร์ชันที่สมบูรณ์ในปี ค.ศ. 2003

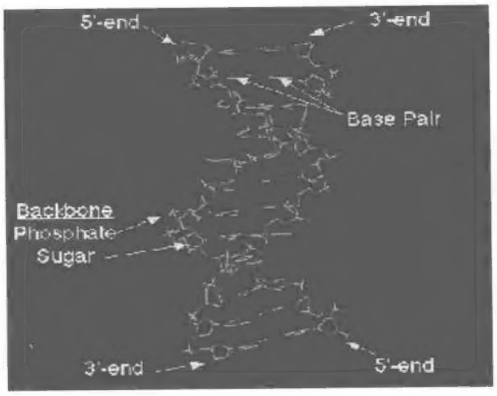
ทำความรู้จักกับความรู้พื้นฐาน DNA

จากการศึกษาสิ่งมีชีวิตพบว่า หน่วยชีวิตเล็ก ๆ เรียกว่า เซลล์ เซลล์เป็นหน่วยหนึ่งที่ทำหน้าที่หรือทำกิจกรรมต่าง ๆ มีการแบ่งเซลล์ เซลล์ใหม่จะได้รับการถ่ายทอดลักษณะทางพันธุกรรมมาจากเซลล์เดิม

ภายในเซลล์ประกอบด้วยส่วนที่อยู่ภายในที่สำคัญเรียกว่า นิวเคลียส ภายในนิวเคลียสมีสารที่สำคัญที่ทำหน้าที่ควบคุมโครงสร้างการทำงานของสิ่งมีชีวิต และถ่ายทอดทางพันธุกรรมเรียกว่า โครโมโซม โครโมโซมมีลักษณะเป็นเส้นยาวขดไปขดมาเต็มไปหมด สำหรับของคนมีโครโมโซมอยู่ 23 คู่ (46 แท่ง) ในสิ่งมีชีวิตต่าง ๆ จะมีจำนวนแท่งโครโมโซมไม่เท่ากัน

แท่งโครโมโซมที่เป็นสายยาวประกอบด้วย กรดนิวคลีอิก ชนิดดีออกซีไรโบ นิวคลีอิก (Deoxyribonuclic acid) เรียกย่อ ๆ ว่า DNA DNA เป็นสารพันธุกรรมที่มีโปรตีนเป็นองค์ประกอบ หน้าที่ของโครโมโซมมีหน้าที่ควบคุมกระบวนการต่าง ๆ ในสิ่งมีชีวิต ควบคุมลักษณะและการถ่ายทอดลักษณะกรรมพันธุ์ต่าง ๆ

หัวใจของลำดับรหัสพันธุกรรมอยู่ที่โครงสร้างระดับโมเลกุลที่จับตัวของสาร DNA ซึ่งประกอบด้วยโครงสร้างทางเคมี 4 แบบ ซึ่งได้แก่ อะดีนีน (Adenine) ไทมีน (Thymine) กัวนีน (Guanine) และไซโตซีน (Cytosine)



รูปที่ 1 ภาพวาดจินตนาการของสายรหัสพันธุกรรมในระดับโมเลกุล
ที่มา : http://www.blc.arizona.edu/

หรือเรียกย่อ ๆ ว่า A T G C ซึ่งสารนี้จะจับตัวกับโมเลกุลของน้ำตาลดีออกซีไรโบสกับโมเลกุลของฟอสเฟส ทำให้ได้โครงสร้างเป็นหน่วยหน่วยหนึ่งเรียกว่า นิวคลีโอไทด์ (ดูรูปที่ 1)

นิวคลีโอไทด์จึงเป็นแกนหลักของ DNA ประกอบเป็นสายยาว และอยู่ร่วมกันเป็นสายคู่ และม้วนพันกันเสมือนเป็นคั่นบันได โดยมีสองคั่นบันไดม้วนซ้อนอีก แต่ละขั้นบันไดที่จับตัวของ DNA จะประกอบด้วยโมเลกุลทางเคมีของหน่วย A, T, G และ C ซึ่งประกอบในนิวคลีโอไทด์ โดยที่แต่ละหน่วยของนิวคลีโอไทด์ที่เป็นโมเลกุลของ A, T, G หรือ C เรียกว่า เบส (Base) การจับกันของเบสเป็นคู่ลำดับที่เรียกว่า เบสแพร (Base pair) โดย A จะอยู่กับ T และ G อยู่กับ C

ลำดับเบสจึงเป็นส่วนสำคัญของโครโมโซม ซึ่งถ้าเรียงลำดับเบสก็จะได้สายอักขระตัวอักษร A, T, G, C ยาวมากมาย ลำดับเบสนี้บางส่วนเท่านั้นที่มีลักษณะสำคัญที่สามารถสั่งการควบคุมการทำงานของสิ่งมีชีวิต และถ่ายทอดออกมาเป็นลักษณะทางพันธุกรรม กลุ่มของลำดับเบสมีลักษณะเป็นโปรตีนที่สร้างฟังก์ชันการทำงานของร่างกาย และเป็นประตูรับคำสั่งเพื่อดำเนิน กรรมวิธีต่าง ๆ จึงเป็นส่วนที่สำคัญที่นักวิทยาศาสตร์ต้องเรียนรู้และเข้าใจว่าลำดับเบสส่วนใดมีฟังก์ชันและหน้าที่อย่างไร

นักวิทยาศาสตร์มีความเพียรพยายามที่จะศึกษาและทำความเข้าใจในลำดับสายรหัส พันธุกรรม จากการศึกษาและเข้าใจโครงสร้างทางโมเลกุล ทำให้สามารถแยกส่วนของ DNA บางส่วนออกมา และทำการหาสายรหัสพันธุกรรมโดยไม่ต้องทำทีเดียวทั้งหมด ลักษณะดังกล่าวเป็นการคัดลอกบางส่วนของยีน โดยเป็นหลักการที่สำคัญของเครื่องมือวิทยาศาสตร์ที่เรียกว่า DNA Sequencing การคัดลอกนี้จะคัดลอกออกมาด้วยความยาวเท่าใดก็ได้ ดังนั้นจึงสามารถใช้เครื่องหาลำดับ DNA นี้ หลาย ๆ เครื่อง ในห้องปฏิบัติการและช่วยกันทำ ซึ่งต้องใช้เวลามาก สำหรับการหาลำดับ DNA ของมนุษย์นั้นได้เริ่มทำมาตั้งแต่ปี 1990 จนถึงปี 2000 ที่บริษัท ซีลีรา ประกาศความสำเร็จ

ข้อมูลของลายพิมพ์รหัสพันธุกรรมที่จัดทำขึ้น จึงเป็นฐานข้อมูลที่จะใช้วิเคราะห์และสร้างความรู้ที่เกี่ยวข้องกับสิ่งมีชีวิต นับเป็นเรื่องที่ท้าทายความสามารถที่จะไขลายพิมพ์ที่สร้างขึ้นจากดีเอ็นเอให้กระจ่าง

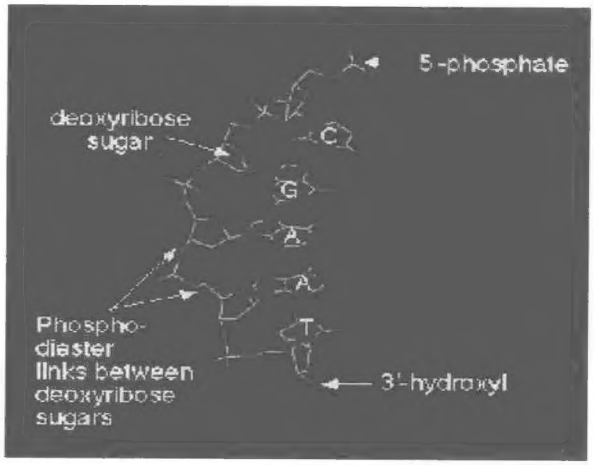
การวิเคราะห์ทางคอมพิวเตอร์

จากที่กล่าวแล้วว่า สายดีเอ็นเอมีโมเลกุลฟอสเฟตและโมเลกุลน้ำตาลไรโบส เป็นแกนหลักที่เรียกว่าแบกโบน คล้ายราวบันไดที่บิดเป็นเกลียว และมีเบสยื่นออกมาจากแกนแบกโบนแต่ละข้างคล้ายขั้นบันได ซึ่งความแตกต่างของแต่ละนิวคลีโอไทด์ขึ้นอยู่กับเบส

เบสที่เป็นส่วนประกอบของดีเอ็นเอมีด้วยกัน 2 ชนิด คือ พิวรีน (Purine) ซึ่งมีโครงสร้างเป็นวงแหวนคู่ และไพริมิดีน (Pyrimidine) ซึ่งมีโครงสร้างเป็นวงแหวนเดี่ยว เบสที่เป็นพิวรีน ได้แก่ อะดินีน (Adenine: A) กับกัวนีน (Guanine: G) ส่วนเบสที่เป็นไพริมิดีน ได้แก่ ไธมีน (Thymine: T) กับไซโตซีน (Cytosine: C) และเบสจะจับกับคาร์บอนที่ตำแหน่ง 1' ของน้ำตาลไรโบส

นักวิทยาศาสตร์ได้นำสายดีเอ็นเอไปตรวจวิเคราะห์ โดยที่สายรหัสของดีเอ็นเอในบุคคลเดียวกันจะมีลายพิมพ์เหมือนกัน ไม่ว่าจะได้จากเซลล์ส่วนใดของร่างกาย เช่น เม็ดเลือดขาว เส้นผม ตัวอสุจิ เป็นต้น ในการนำสายดีเอ็นเอไปใช้ในงานต่าง ๆ เช่น พิสูจน์ความเป็นพ่อแม่ลูก พิสูจน์ในคดีฆาตกรรม นักวิทยาศาสตร์จะนำสายดีเอ็นเอไปตัดด้วยเอนไซม์จำเพาะชนิดที่กำหนดไว้แล้วตามช่วงของสายดีเอ็นเอที่ต้องการพิสูจน์ และทำลายพิมพ์ดีเอ็นเอออกมา จะเห็นว่าในการเก็บข้อมูลเกี่ยวกับดีเอ็นเอ หากเก็บเป็นภาพลายพิมพ์ดีเอ็นเอของสิ่งมีชีวิตทุกชนิดบนโลกจะได้ฐานข้อมูลที่มีขนาดใหญ่มาก ดังนั้นการเก็บข้อมูลดีเอ็นเอ จึงเก็บในรูปของตัวอักษรเรียงต่อกันเป็นลำดับเบส ซึ่งมีตัวอักษรอยู่ 4 ตัว คือ A, G, C และ T เป็นชนิดของเบสดังที่กล่าวมาแล้วข้างต้น และตัวอักษร N แทนลำดับเบสที่ไม่รู้จัก ซึ่งการนำลายพิมพ์ดีเอ็นเอที่ได้ไปหาลำดับเบสนั้นมีลักษณะดังรูปที่ 2 ซึ่งจะได้ลำดับเบสของดีเอ็นเอคือ CGAAT ถึงแม้ว่าฐานข้อมูลจะเก็บอยู่ในรูปของลำดับเบสก็ตาม แต่ก็ยังมีขนาดใหญ่มากอยู่นั่นเอง

ฐานข้อมูลลำดับเบสของดีเอ็นเอนั้นได้รับการเก็บรวบรวมมาจากเครื่องหาลำดับเบสของดีเอ็นเอ และนำมาเก็บรวมไว้เป็นฐานข้อมูลหลักที่องค์กร National Center for Biotechnology (NCBI) หรือ GenBank โดยฐานข้อมูลเหล่านี้แบ่งเป็นหมวดหมู่ย่อย ๆ ตามลำดับเบสที่ได้มาจากสิ่งมีชีวิตต่างชนิด และชิ้นส่วนของอวัยวะของสิ่งมีชีวิตที่ต่างกัน เช่น แบคทีเรีย ยีสต์ หนู แมลงหวี่ มนุษย์ และไมโตคอนเดรีย (Mitochondria) ของสัตว์ ชนิดต่าง ๆ เป็นต้น นับว่าเป็นฐานข้อมูลที่มีความหลากหลายและมีขนาดใหญ่มาก นอกจากนี้มีหน่วยงานต่าง ๆ จำนวนมาก เช่น มหาวิทยาลัยเกษตรศาสตร์ ได้คัดลอกจัดทำระบบฐานข้อมูล และเรียกว่า biomirror (http://bio-mirror.ku.ac.th)



รูปที่ 2 ตัวอย่างของ DNA Backbone: 5'-d (CGAAT) (ที่มา : http://www.blc.arizona.edu)

การเก็บข้อมูลในฐานข้อมูลลำดับเบสจะเก็บอยู่ในรูปแบบของ FASTA ซึ่งได้แสดงไว้ดังรูปที่ 3 โดยมีเครื่องหมาย ">" เป็นตัวแยกชนิดของลำดับเบสของดีเอ็นเอ และหลังเครื่องหมาย ">" ก่อนข้อมูลลำดับเบส จะเป็นข้อมูลของชนิดของลำดับเบสนั้น ๆ โดยข้อมูลเหล่านั้นจะบอกถึง

- แหล่งที่ได้ลำดับเบส (gb ย่อมาจาก Genbank)
- ชนิดของลำดับเบส (Escherichia coli, tesA, ybbA genes)
- ตำแหน่งคู่เบสของลำดับเบส (from base s 510705 to 522297 of complete genome)

ส่วนข้อมูลลำดับเบสจะเป็นตัวอักษรของเบสเรียงต่อกันเป็นข้อความ ซึ่งชนิดของลำดับเบสที่ต่างกันก็จะมีข้อมูลลำดับเบสที่ต่างกันด้วย ดังตัวอย่างในรูปที่ 3

การดำเนินการทางด้านคอมพิวเตอร์ จึงเสมือนการค้นหาข้อความภายในว่ามีความหมายอะไรบ้าง นักวิทยาศาสตร์ต้องการอ่านรหัสเหล่านี้ออกมาในเชิงความหมาย ซึ่งแน่นอนต้องมีการเปรียบเทียบระหว่างสายพันธุ์ ระหว่างกลุ่มของสิ่งมีชีวิตอื่น เพื่อจะดูความหมายการจัดเรียงตัวเหล่านี้

รูปที่ 3 ตัวอย่างข้อมูลลำดับเบสของดีเอ็นเอในฐานข้อมูลลำดับเบส (ที่มา : http://www.ncbi.nlm.nih.gov)

งานที่ต้องดำเนินการต่อ

ปัจจุบันมีเครื่องวิเคราะห์สายยีน หรืออุปกรณ์ที่ทำทำหน้าที่หาลำดับเบสของดีเอ็นเอ มีกระจายอยู่ทั่วโลก หน้าที่หลักคือการช่วยกันดำเนินการเพื่อหาลำดับเบสของสิ่งมีชีวิตที่สนใจ และรวมกันเป็นฐานข้อมูลเพื่อไว้ใช้ศึกษา

การที่มีข้อมูลอยู่ในฐานข้อมูลและต้องใช้นักวิทยาศาสตร์ นักชีววิทยา นักคอมพิวเตอร์ มาร่วมกันศึกษาและวิเคราะห์ โดยเฉพาะอย่างยิ่งจากตัวอย่างของรหัสพันธุกรรมของมนุษย์ ซึ่งเป็นลำดับเบส ATGC จำนวนมาก ประมาณ 3 พันล้าน เชื่อกันว่าจะมีส่วนสำคัญของยีนที่มีบทบาทสำคัญต่อมนุษย์อยู่ประมาณ 25,000-150,000 ซึ่งเป็นส่วนที่น่าสนใจและต้องศึกษาให้รู้ได้ ข้อมูลส่วนนี้อยู่ใน 3 พันล้านที่ต้องหามาให้ได้

การศึกษาในเรื่องรหัสพันธุกรรมนี้ ยังต้องศึกษาข้อมูลของรหัสพันธุกรรมของสายพันธุ์อื่นด้วย เพื่อจะทำให้ได้ความรู้ต่าง ๆ อีกมากมายที่จะไขปัญหาของสิ่งมีชีวิต บริษัทซีลีรา จึงได้ทำการเก็บข้อมูลรหัสพันธุกรรมของแมลงวันผลไม้ โดยทำการทดลองหาลำดับเบสทางพันธุกรรมประมาณ 120 ล้านเบสแพร์ เพื่อทำการศึกษาทดลองต่าง ๆ และก่อนที่จะทำการศึกษาและค้นหาความลับที่เกี่ยวกับรหัสพันธุกรรมของมนุษย์ ซีลีราและกลุ่มบริษัทที่ทำการศึกษาร่วมได้ทำการหาสายรหัสพันธุกรรมของหนู และได้ดำเนินการจนได้รหัสทั้งหมดของหนูในขั้นต้น ด้วยขนาดที่ใกล้เคียงกับมนุษย์คือ ประมาณ 3 พันล้านเบสแพร์ ซีลีร่า ได้เก็บรวบรวมข้อมูลและเริ่มต้นทำการวิเคราะห์ค้นหาสิ่งมหัศจรรย์ต่าง ๆ ที่อยู่ในการดำเนินการของสิ่งมีชีวิต

งานการอ่านสายรหัส และการแปลความหมายของสายรหัส เป็นงานที่ยิ่งใหญ่มาก เป็นงานที่เกี่ยวข้องกับข้อมูลที่มีขนาดใหญ่มาก และเป็นงานที่ยิ่งใหญ่ตั้งแต่อดีตที่มนุษย์เคยดำเนินการมา

โดยปกติเครื่องหาลำดับเบส (DNA Sequencing) หนึ่งเครื่อง จะวิเคราะห์หาลำดับเบสของดีเอ็นเอได้ประมาณ 330,000 เบส โครงการจีโนมมนุษย์ หรือ International Human Genome จึงต้องใช้วิธีการร่วมมือกันโดยให้นักวิจัยและศูนย์การทดลองแต่ละแห่งช่วยกันทำและนำข้อมูลมาเก็บรวบรวมไว้ภายใต้โครงการนี้ ฐานใหญ่สุดอยู่ที่ Sanger Center ในประเทศอังกฤษหรือที่สถาบันวิจัยไบโอเมดิคอลที่ Whitehead มลรัฐแมสสาชูเซส มีเครื่องหาลำดับเบสอยู่แห่งละประมาณ 100 เครื่อง สำหรับที่บริษัทซีลีร่าเพียงแห่งเดียวมีเครื่องลำดับเบสอยู่ถึง 300 เครื่อง และมีความสามารถหาลำดับเบสได้ถึง 2 พันล้านเบสในเวลาหนึ่งเดือน ซึ่งมีขีดความสามารถสูงสุดตั้งแต่ที่โครงการเคยดำเนินการมา

อย่างไรก็ตามด้วยเทคนิคทางด้านการจัดการและระบบคอมพิวเตอร์ที่ใช้ในการจัดการข้อมูล ทำให้การดำเนินการต่าง ๆ เป็นไปแบบอัตโนมัติมากขึ้น และมีความน่าเชื่อถือได้สูงขึ้น

เทคนิคทางคอมพิวเตอร์เป็นเรื่องที่มีความสำคัญ

การที่บริษัทซีลีร่าทำงานต่าง ๆ ได้รวดเร็วขึ้น ซีลีร่ากล่าวว่าเกิดจากขีดความสามารถของคอมพิวเตอร์ที่ใช้งานขณะนี้ทำงานได้ดีขึ้น ที่ศูนย์เก็บข้อมูลและวิเคราะห์ของซีลีร่าใช้ซูเปอร์คอมพิวเตอร์ที่ใช้ซีพียูของคอมแพ็คอัลฟ่า ที่ทำงานได้เร็วมากกว่า 250 เครื่องต่อกันเป็นเครือข่าย และยังมีฐานข้อมูลขนาด 4.5 เทราไบต์ นอกจากนี้ยังเชื่อมโยงกับเครื่องคอมพิวเตอร์แบบซูเปอร์คอมพิวเตอร์ที่ใช้ซีพียูอัลฟ่าอีกกว่า 800 เครื่องในที่ต่าง ๆ และมีที่เก็บข้อมูลรวมกันกว่า 50 เทอราไบต์

เทคนิคการคำนวณที่สำคัญคือ การตัดแบ่งส่วนของสาย ATGC ออกเป็นชิ้นเล็ก ๆ การค้นหาตำแหน่งในสาย ATGC การเปรียบเทียบข้อมูล การกระทำหลายอย่างกระทำในลักษณะความคล้ายหรือการกระทำที่เรียกว่า Similar Matching อัลกอริทึม ในการคำนวณบนฐานข้อมูลขนาดใหญ่นี้ ต้องใช้เวลานานและยากที่จะหาคำตอบได้รวดเร็ว ขณะเดียวกันต้องใช้หน่วยความจำของคอมพิวเตอร์มาก และยังต้องใช้ที่เก็บข้อมูลหรือฐานข้อมูลใหญ่เช่นเดียวกัน

ลองนึกถึงสายสตริง (String) ที่ใช้ในการแทนข้อมูลในระบบข้อมูลที่มีขนาดยาวมาก การค้นหาตำแหน่งโดยนำสตริงอีกตัวหนึ่งมาเปรียบเทียบการดำเนินการเช่นนี้จำเป็นต้องหาวิธีการเพื่อให้กระทำได้เร็วขึ้น เมื่อความยาวของสตริงเพิ่มขึ้น ระยะเวลาในการคำนวณก็จะมากตามและเพิ่มขึ้นอย่างมากมาย

รหัสพันธุกรรมจึงเป็นศาสตร์แขนงหนึ่งที่ต้องอาศัยการดำเนินการด้วยคอมพิวเตอร์และเป็นเรื่องที่สำคัญต่อทุกวงการ โดยเฉพาะวงการคอมพิวเตอร์ที่ต้องพบกับการจัดการฐานข้อมูลขนาดใหญ่มาก การคำนวณที่ต้องใช้กำลังคอมพิวเตอร์มากมาย ตลอดจนการบริหารจัดการข้อมูลในลักษณะช่วยกันทำช่วยกันใช้ โดยมีเครือข่ายคอมพิวเตอร์เป็นตัวเชื่อมประสาน

รหัสของสิ่งมีชีวิตจึงเป็นข้อมูลข่าวสารที่จะบอกความเร้นลับของสิ่งมีชีวิต เป้าหมายของการศึกษาทางด้านไบโออินฟอร์มาติก จึงอยู่ที่การค้นหาความรู้ในเรื่อง ต่าง ๆ ที่จะเป็นประโยชน์ต่อการดำรงเผ่าพันธุ์มนุษย์ แน่นอนที่สุดการดำเนินการหลายอย่างย่อมมีผลประโยชน์ทางการค้าแอบแฝงอยู่ด้วย ดังจะเห็นได้จากการทุ่มลงทุนมหาศาลจากการคิดค้น และแสวงหารหัสพันธุกรรม ซึ่งเป็นเรื่องที่ถือว่าเป็นการท้าทายอย่างยิ่งยวด

เอกสารอ้างอิง

1. Samuel K. Moore "Understanding the human genome" IEEE Spectrum Vol.37 No.11, Nov. 2000.